

Exploratory Factor Analysis

Daniel B. Wright and Daniella K. Villalba

Florida International University

Chapter for Breakwell et al. (2011) *Research Methods*

Aims/Learning Objectives

1. To learn what a latent variable is and why exploratory factor analysis (EFA) is used.
2. To be able to interpret computer output for these techniques.
3. To learn about extensions to EFA.

Web page: <http://www2.fiu.edu/~dwright/EFA/>

In 1904 Charles Spearman published one of the most important papers in the history of psychology. This landmark paper made separate contributions for methodology, cognition, and statistics. The methodological contribution was to introduce correlational psychology to complement the dominant experimental approaches of the day of Ebbinghaus, Fenchner, etc. Spearman said the correlational psychology approach was necessary to shed "real light upon the human soul, unlock the eternal antinomy of Free Will, [and] ... reveal the inward nature of Time and Space" (1904, p. 203). The contribution to cognition was a description of the notion of general intelligence, or *g*. The statistical contribution was to show a way in which to estimate this general intelligence variable even though it was not directly measured. While his method built upon earlier work by people such as Galton and Pearson, Spearman's procedure is generally regarded as the beginning of latent variable modeling. Spearman's *g* is a latent variable because it is not directly observed. This contribution is the focus of this chapter. Estimating unobserved or latent variables is critical in all areas of psychology (but particularly in psychometrics, see Chapter *). We often measure several observed or manifest variables in order to tap into some construct which is not directly observed.

Researchers use latent variable models when they believe there are just a small number of hidden variables that cause the associations among the observed variables. In the first section we present these relationships with diagrams and show how these diagrams can be translated into equations. We describe some of the general issues that are important for understanding latent variable models. Detailed mathematics of latent variable models is beyond the scope of this book (but covered in most multivariate statistics books aimed at psychologists and we provide a primer for matrices in an appendix), but thanks to computers and statistical packages these models can be calculated by anyone with good conceptual knowledge of latent variables. In the second section we show examples of latent variable models and how to interpret the output from a couple of popular statistics programs. In the third section, we mention two ways to extend the basic latent variable model. We end with a summary and some recommendations.

The focus of this chapter is on one type of latent variable model called exploratory factor analysis, often abbreviated EFA. Within EFA the latent variables are called factors. Other latent variable models and an alternative to EFA called principal components analysis (or PCA) are briefly discussed at the end of the chapter.

Warning 1: The observed variables need to be correlated among themselves. After doing exploratory data analysis (EDA, Chapter *), look at the pairwise correlations (Chapter *), and create scatter plots. If the correlations are all small (i.e., less than about .3 in magnitude) then it does not make sense to hypothesize that a single latent variable, or a small set of latent variables, is influencing the observed variables.

Warning 2: There is a little more mathematics in this chapter than in most of the other chapters. That is inevitable given the subject matter.

1. What is a Latent Variable

Latent means hidden, but still exerting some effect. For example, we research something called "social avoidance" (and its role in the eternal antimony of Free Will). People who have a lot of social avoidance try to avoid social situations. When forced to be in social situations they do not engage with other people as much as those who are less socially avoidant. At one end of the scale would be hermits and at the other end would be the social butterflies. You can think where different people, for example, Theodore Kaczynski (the Unabomber) and Britney Spears (a mouseketeer), might be on this scale. Or think whether people who play team sports tend to be less social avoidant than those playing individual sports. The next party you are at (going to a party makes you less social avoidant than Kaczynski) try observing behaviors that you think might be consistent with either high social avoidance (e.g., doing the dishes alone in the kitchen, drinking in the corner) or low social avoidance (e.g., handing out phone number to everyone, talking when no one is listening). The theory we use states that this unobserved construct, social avoidance, *causes* people to behave in ways that we can observe.

Psychologists often ask people for their responses on pen-and-paper questionnaires (Chapter *). Suppose we asked people to respond to the questions in Table 1. For each of these we expect that having a lot of social avoidance will lead people to circle higher numbers. None of these directly measure social avoidance, but we expect a socially avoidant person to be more likely to sit by a lake in Greenland, watching TV, and waiting for mom to call, than going clubbing.

Insert Table 1 about here

Latent Variables in Diagrams and Equations

Figure 1 shows how we imagine the relationship between social avoidance and these responses. It is convention to show latent variables with ellipses and observed variables with rectangles. In this example, social avoidance is a latent variable and the responses for each of the behaviors are observed variables. There is an arrow going from social avoidance to each of the observed variables. According to this model responses to each of these questions are influenced by the latent variable "social avoidance". There are also individual circles to the right of each behavior with an arrow going to each of the observed behaviors. These are called the *item-*

specific errors (also called item-unique errors), but they are better thought of as just idiosyncratic variation for that question. For example, a social butterfly might really like tortilla chips and live far from the stadium, so might opt to watch football on TV rather than go to the game. This is an example of item specific error. Often researchers do not draw these item-specific error variables in their diagrams, but it is worth remembering that each observed variables has its own error term.

Responses to each observed variable are affected by the social avoidance latent variable and an item-specific error. Let's use the following notation: $Greenland_i$ to refer to the first observed variable, $SocAvoid_i$ to refer to the social avoidance latent variable, and $e1_i$ to be the item-specific error variable for the first question. Notice how we have included the subscript i on each of these to show that each individual person can have different values on these. We can write this as:

$$Greenland_i = \alpha1 SocAvoid_i + e1_i$$

where $\alpha1$ (read: alpha one) is the strength of the influence of the latent variable on the observed variable, also referred to as a *loading*. This equation is similar to the regression equations in Chapter *, except there is no intercept. This is because the mechanics of solving latent variable models re-scales the variables (if you'd like, imagine $Greenland_i$ has been standardized to have a mean of zero and a standard deviation of one). The scaling of the variables also means that $\alpha1$ is between -1 and 1 and can be interpreted in a similar way to Pearson's correlation r between the observed variable and the latent variable (for this simple model). Similar equations can be used for the remaining observed variables:

$$\begin{aligned} Football_i &= \alpha2 SocAvoid_i + e2_i \\ Lake_i &= \alpha3 SocAvoid_i + e3_i \\ Phone_i &= \alpha4 SocAvoid_i + e4_i \end{aligned}$$

Complex statistical techniques provide estimates for the α s.

Insert Figure 1 about here

Human behavior is complex and is governed by more than just one underlying latent variable. Social avoidance is usually thought of as one of the two components of social anxiety. The other component is a fear of negative evaluation. Suppose we asked a sample of people whether in the previous week they had done each of 8 behaviors that clinicians describe as typical for social anxiety. It might be that social avoidance increases the chances that you will do some behaviors ($x1, x2, x3$ in Figure 2), fear of negative evaluation increases others ($x6, x7, x8$), and both factors are predictive of doing other behaviors ($x4, x5$).

Insert Figure 2 about here

Figure 2 can be translated into a system of equations.

$$\begin{aligned} x1_i &= \alpha21 SocAvoid_i + & e1_i \\ x2_i &= \alpha21 SocAvoid_i + & e2_i \\ x3_i &= \alpha31 SocAvoid_i + & e3_i \end{aligned}$$

$$\begin{aligned}
 x_{4i} &= \alpha_{41} \text{ SocAvoid}_i + \alpha_{42} \text{ Fear}_i + e_{4i} \\
 x_{5i} &= \alpha_{51} \text{ SocAvoid}_i + \alpha_{52} \text{ Fear}_i + e_{5i} \\
 x_{6i} &= \alpha_{62} \text{ Fear}_i + e_{6i} \\
 x_{7i} &= \alpha_{72} \text{ Fear}_i + e_{7i} \\
 x_{8i} &= \alpha_{82} \text{ Fear}_i + e_{8i}
 \end{aligned}$$

Most computer programs output the α values in tables like these:

$$\begin{array}{c}
 \text{Loadings} = \mathbf{L} = \boldsymbol{\alpha} = \\
 \left[\begin{array}{cc}
 \alpha_{11} & \alpha_{12} \\
 \alpha_{21} & \alpha_{22} \\
 \alpha_{31} & \alpha_{32} \\
 \alpha_{41} & \alpha_{42} \\
 \alpha_{51} & \alpha_{52} \\
 \alpha_{61} & \alpha_{62} \\
 \alpha_{71} & \alpha_{72} \\
 \alpha_{81} & \alpha_{82}
 \end{array} \right]
 \end{array}
 \quad \text{or like} \quad
 \begin{array}{c}
 \mathbf{SA} \quad \mathbf{FNE} \\
 \left[\begin{array}{cc}
 \alpha_{11} & \\
 \alpha_{21} & \\
 \alpha_{31} & \\
 \alpha_{41} & \alpha_{42} \\
 \alpha_{51} & \alpha_{52} \\
 & \alpha_{62} \\
 & \alpha_{72} \\
 & \alpha_{82}
 \end{array} \right]
 \end{array}$$

This is called the $\boldsymbol{\alpha}$ or loadings matrix (see appendix for discussion of matrices). The α_{31} , for example, corresponds to the arrow in Figure 2 connecting the third observed variable with the first latent variable. We have written this table in two ways. In the first all the α values are written, even small ones (we have placed these in grey, most computer packages do not do this). In the second we have just left out α values that are small. Many computer programs allow the option not to print small α values. In terms of the diagram, it means that there is not an arrow connecting that latent variable to that behavior.

We have added the labels *SA*, for social avoidance, and *FNE*, for fear of negative evaluation, above the columns in the equation above. We looked at which variables loaded on the same construct. We saw the first 5 all loaded on the same construct and our theory suggests this could be social avoidance. The final 5 all load together so we felt this could be fear of negative evaluation. The naming of these factors is subjective, and is discussed in more detail later. If we assume the first factor is social avoidance and the second is fear of negative evaluation, then we can say, for example, that social avoidance does not influence the likelihood of doing behavior #6 because α_{61} is small. There is a bit of circularity.

1. We defined the first factor "social avoidance" because of the variables that load and do not load on it.
2. Then we say "social avoidance" does not relate to one of the variables that we had used to inform how we defined it.

Care is needed when interpreting the relationship between individual observed variables and the psychological meaning of factors.

There are several different methods for estimating the α values. Tabachnick and Fidell (2007) discuss seven different methods available in the packages SAS and SPSS/PASW. The package CEFA, discussed on our web page, offers 8. The method used in our examples is called maximum likelihood, but all the methods have their purposes (and usually give similar results). It is often worthwhile trying a few methods and seeing if you get similar results.

Four Principles of Science and Statistics

1. Scientists strive for simple theories. This is called *Occam's razor* or the law of parsimony.
2. Scientists strive for theories that account for a lot of the variation in their data.
3. Analyses should reveal exciting new findings that would not be available by looking at the data without the analyses.
4. Scientists strive for theories that tell a good story (Abelson, 1995).

Deciding the number of latent variables

Scientists like simple theories (Principle #1 above) and like theories that account for a lot of their data (Principle #2 from above). Unfortunately, there is usually a trade-off between the simplicity of the theory and how much of the data is accounted for. One measure of simplicity is the number of latent variables used to account for the data. The more latent variables there are; the more complex the model is. There are a few ways to estimate how well the model accounts for the data, but the most common way is to see how much of the variation in the data is accounted for by the model. The statistical packages calculate something called the *communality* for each of the observed variables.¹ This is the sum of the squared α values for that variable, even the small ones which do not warrant arrows in Figures 1 and 2. The communality for question #1 would be:

$$c1 = \alpha_{11}^2 + \alpha_{12}^2$$

This value will usually be between 0 and 1.² The communalities of all the observed variables can be added together to create the total variance accounted for by all the latent variables. This is the amount of the variation in the original variables accounted for by the latent variables in units of *eigenvalues*. The word "eigenvalues" relates to the mathematics of factor analysis. An eigenvalue value is a unit of measurement for the proportion of variation accounted for by the latent variables. If this total is divided by the number of observed variables you get the proportion of the total variation accounted for. Like meters and yards are both units of measurement of length, eigenvalue is just like using a ruler with different numbers. If there are 20 observed variables, then 1 eigenvalue of variation corresponds to 1/20th or 5% of the total variation. All that is necessary to know here is that one eigenvalue is the same as the average variation in each observed variable.

¹ Some packages print the *uniqueness* of a variable which is 1 minus the communality.

² Occasionally you do get communalities outside of this range. This usually means there is a problem, like some particularly high correlations. If some variables are correlated above .8 or so, try either removing one of the variables or calculating the mean of them and using the mean of these rather than both of the original variables.

Principles #1 and #2 from the box above often compete. To examine this trade-off between simplicity and how much of the observed data is accounted for researchers plot the two together. For factor analysis this means plotting the number of latent variables (complexity) on the x-axis and the proportion of variation accounted for on the y-axis. Researchers usually plot how much additional variation is accounted for each time a latent variable is added to the model. This is called a *scree* plot. In geological terms the scree is the rubble that is at the bottom of a steep mountain. It is touching the mountain but is just loose rubble. Cattell (1966) used this metaphor to describe where the statistical model starts accounting to random variation (i.e., the loose rubble). You want the model to account for the mountain, not the loose rubble. He said to choose the model with the most latent variables that is not attempting to account for the loose rubble. This is a great metaphor. Metaphors are a powerful tool both to help construct models in science and to help explain them to others. Figure 3 shows a scree plot superimposed onto its geological namesake. There appears to be two latent variables that are accounting for the structure. Some people use a biological metaphor and say to choose the model at the *elbow* in the line.

Insert Figure 3 about here

Some statistical packages will print a dotted horizontal line at the level equivalent to 1 eigenvalue. This is called the Kaiser (1960) criterion (sometimes labeled Kaiser-Guttman). Kaiser (1960) argued that if adding an additional latent variable contributed less than one eigenvalue then it may not worth including. Textbooks disagree whether the cutoff should be exactly 1 eigenvalue, or slightly more or less.

Another way to compare models is with a significance test. Several of the computer programs will print a χ^2 value for how far off the model is from the observed values. This is like the residual sum of squares in ANOVA, but a χ^2 value is produced (if the computer prints more than one χ^2 value, use the one labeled "likelihood ratio χ^2 "). Two models can be compared if one is nested within the other, meaning that the two are the same except that one of them has something extra. Here the extra would be that it allows an extra latent variable. The difference between the two χ^2 values is a measure of how much more variation is accounted for by the more complex model. The more complex model will almost always account for more of the data. Significance testing helps us to explore whether the additional variation accounted for is greater than chance. The difference can be looked up on a χ^2 table. The degrees of freedom are equal to the number of extra parameters estimated. If the only difference is the inclusion of one extra latent variable, there will be $k - 1$ more degrees of freedom where k is the number of observed variables. This is because a loading is calculated for each observed variable for the new latent variable. The difference is not k degrees of freedom because of the constraint placed on the loading values that the sum of their squared values is one. We will show an example of this in the second section of this chapter.

Finally, it is worth seeing how similar the correlation matrix implied by the model is to the observed correlation matrix. EFA works just on the correlation matrix so this is an important way to see, overall, if the model works, and also to see if there are any particular correlations that the model fails to predict. The method to calculate the implied correlation is to multiply the loadings which connect paths between the items. For example, items 4 and 5 in Figure 2 each have arrows connecting them from both latent variables. The correlation due to the first factor is α_{41} multiplied by α_{51} and due to second factor is α_{42} multiplied by α_{52} . Their sum is: $\alpha_{41} \cdot \alpha_{51}$

+ $\alpha_2 \cdot \alpha_5$. For some complex models it is necessary to trace all the arrows for every possible implied correlation. Fortunately, the loadings matrix can be multiplied by its transpose (the matrix is flipped on its side, see appendix) to create the implied correlations. One thing to note is that this can be done with all the loadings or just those above a certain amount. In the first example we show both methods. When you get the implied correlation matrix, it is useful to subtract it from the observed correlation matrix. This will help you to locate which correlations the model is not able to re-produce well.

Some people believe that there is debate about which *single* method should be used to decide how many latent variables to have. In reality, no (respectable) researcher would use only one method. It is worth trying a few. Overall, while the scree plot is very subjective, we find it the most useful.

Rotation

Once you have decided how many latent variables to have, then you have to decide what they are. If there is only one latent variable, then look at the loadings and try to come up with a name that summarizes all the observed variables which have high loadings. **When you have a single latent variable you do not use rotation.**

Suppose that you decide that there are two latent variables. The initial solution that the computer outputs gives the loadings, the α s, and this allows you to plot each of the observed variables. These are called the unrotated loadings. Here is what they might look like for 10 observed variables:

$$\text{unrotated loadings} = \begin{bmatrix} -.75 & -.42 \\ -.29 & -.22 \\ -.01 & .05 \\ .32 & .11 \\ .78 & .43 \\ .44 & -.59 \\ .09 & -.30 \\ .09 & -.01 \\ -.20 & .33 \\ -.35 & .62 \end{bmatrix} = \begin{bmatrix} -.75 & -.42 \\ -.29 & -.22 \\ -.01 & .05 \\ .32 & .11 \\ .78 & .43 \\ .44 & -.59 \\ .09 & -.30 \\ .09 & -.01 \\ -.20 & .33 \\ -.35 & .62 \end{bmatrix}$$

We wrote these twice. The second matrix shows loadings less than .20 in magnitude in grey. Figure 4a shows a scatter plot of the variables. Figure 4b shows this in a causal diagram. If we tried to describe these latent variables it would be difficult. The first latent variable is positively related to observed variables 4, 5, and 6, and negatively related to variables 1, 2, 9, and 10. The second latent variable is positively related to variables 5, 9, and 10, and would be negatively related to variables 1, 2, 6, and 7. This is complex. Ideally you would want your latent variables not to relate to a lot of the same variables. Here, variables 1, 2, 5, 6, 9, and 10 all are influenced by both latent variables. It would be difficult to tell a good story with this solution.

Insert Figures 4a and 4b about here

When you conduct a factor analysis, the computer helps you to decide how many latent variables are necessary. It allows you to draw a plot like Figure 4a. However, its choice of where to put the axes is somewhat arbitrary. If you rotate the axes around by rotating this book, this does not affect the number of latent variables, the total amount of variation accounted for by the

model, or the communalities of any of the observed variables. Therefore, the data (or the axes) can be rotated to see if a solution can be found that is easier to explain. This sounds like cheating, but it is not.

Why would we want to do rotation? The answer is that it can help us to better understand the loadings. Recall from Chapter * that one goal of creating a good psychometric test is having items that load on one and only one construct. Consider the point in the lower-left hand side of the left panel of Figure 4. It has large negative values for both its x value (-.75) and its y value (-.42). We would prefer measures which have large values (positive or negative) on only one of the variables. This makes them better indicators of that latent variable.

Looking at Figure 4a we can think what the values would be like if we used different axes. We can rotate the axes until most of the observed variables load highly on only one of the latent variables, but not on the other. Ideally each behavior would load on one and only one latent variable. In practice, observed variables will likely have at least a small loading on every latent variable, but if the loadings are below about .2 we usually treat them as if the latent variable does not influence that observed variable. We could guess what rotation will make variables tend to load on only one factor, but clever statisticians have developed several methods to choose how much to rotate the points. And clever computer scientists have written these methods into statistics packages. There are two main types of rotation: orthogonal (uncorrelated) and non-orthogonal (correlated). Orthogonal methods are more common and the resulting variables are more easily included in other procedures. Tabachnick and Fidell (2007) describe ten different methods available in some statistical packages (and there are more). We will consider only the most used method, called *varimax*. It tries to make each factor either load highly on an observed variable or not load on it at all. If we did this with the above data the computer would tell us that a rotation of 25.91° was what the statistician's clever function recommends.

To rotate the data 25.91° we use the following *rotation matrix* (see any trigonometry book for cosine and sine, abbreviate cos and sin, and see the appendix for how to multiply matrices):

$$\begin{bmatrix} \cos 25.91^\circ & -\sin 25.91^\circ \\ \sin 25.91^\circ & \cos 25.91^\circ \end{bmatrix} = \begin{bmatrix} .90 & -.44 \\ .44 & .90 \end{bmatrix}$$

To calculate the degree of rotation from the numbers you need to use the inverse sine (the arc-sine) or the inverse cosine (the arc-cosine) functions ($\pi \approx 3.14$):

$$\text{arc-cosine}(.90) * 180/\pi = \text{arc-sine}(.44) * 180/\pi \approx 26^\circ \text{ (approx. because of roundings)}$$

Mathematically, the rotation is done by multiplying the unrotated loadings by the rotation matrix:

$$\begin{array}{cc}
 \textit{unrotated} & \textit{varimax} \\
 \textit{loadings} & \textit{rotation} \\
 \begin{bmatrix} -.75 & -.42 \\ -.29 & -.22 \\ -.01 & .05 \\ .32 & .11 \\ .78 & .43 \\ .44 & -.59 \\ .09 & -.30 \\ .09 & -.01 \\ -.20 & .33 \\ -.35 & .62 \end{bmatrix} & \begin{bmatrix} .90 & -.44 \\ .44 & .90 \end{bmatrix} = \\
 & \textit{rotated} \\
 & \textit{loadings} \\
 \begin{bmatrix} -.86 & -.05 \\ -.35 & -.08 \\ .01 & .05 \\ .34 & -.04 \\ .89 & .04 \\ .14 & -.72 \\ -.05 & -.31 \\ .07 & -.05 \\ -.03 & .39 \\ -.04 & .71 \end{bmatrix} & = \\
 & \textit{rotated} \\
 & \textit{loadings} \\
 \begin{bmatrix} -.86 & .05 \\ -.35 & -.08 \\ .01 & .05 \\ .34 & -.04 \\ .89 & .04 \\ .14 & -.72 \\ -.05 & -.31 \\ .07 & -.05 \\ -.03 & .39 \\ .04 & .71 \end{bmatrix}
 \end{array}$$

Figure 5b shows the rotated version of Figure 4b. There were 14 arrows connecting latent variables to observed variables (not counting the item specific error latent variables) in Figure 4b. There are only 8 in Figure 5. Further, now none of the observed variables are influenced by both of the latent variables. The theory suggested from Figure 5b is simpler than the model suggested by Figure 4b. Latent variable 1 influences responses to observed variables 1, 2, 4, and 5, while latent variable 2 influences responses to observed variables 6, 7, 9, and 10.

Insert Figures 5a and 5b about here

The main non-orthogonal rotations are *promax* and *direct oblimin*. Both of these allow you to vary how correlated the latent variables are. Non-orthogonal rotation means the x and the y axes are rotated different amounts. This means that the resulting factors can be difficult to use in subsequent statistical methods and the mathematics is more difficult.

Recommendation: The most common rotation method is varimax, but other methods are available with most statistics packages. It can be valuable to try several rotations and use whichever rotation makes it easiest to interpret the loadings. The purpose of rotation is to make it easier to interpret the loadings, so this is *not* cheating! If you think the latent variables are likely to be correlated, you can use the non-orthogonal rotation methods (and try different amounts of correlation). Most EFA uses orthogonal rotation.

When there are two latent variables it is fairly easy to conceive of rotating a two dimensional piece of paper in order to visualize rotation in two dimensions. When there are three latent variables it is more difficult to conceptualize rotating a three dimensional space and it is extremely difficult to conceive of rotating a four dimensional object, but the actual algorithms used for rotation have no difficulty with it. It is best just to believe that the rotation methods work regardless of the number of dimensions.

If you have 3 latent variables, the rotation matrix is a 3x3 matrix, and so on for models with more latent variables. When there are more than 2 latent variables the mathematics gets

more complex and you cannot make scatter plots like in Figures 4a and 5a. But you can and should make the diagrams in Figures 4b and 5b.

Naming the Latent Variables

One of the most important steps in factor analysis is naming the latent variables. You need to look at the items with high loadings on each latent variable and decide a label that summarizes these. Hopefully the rotation will produce a model which makes this easier. This is an important step because you will be using these labels to describe your theories. The labels will take on a life of their own. Imagine how labeling a latent variable "anti-social personality type" will affect how you describe your theory compared with the label "social avoidance". It is also important that all the items that load on the factor appear related to the chosen label, and that all the items that appear related to the chosen label load on the factor.

Labeling latent variables is not something that statisticians have any special knowledge with which to help researchers (though researchers often ask). Both knowledge about the area and thinking pragmatically about what label will be best to convey your findings are necessary for naming variables. And a bit of pizzazz doesn't hurt. One method to check the validity of your labels is to write on little slips of paper all the questions that you asked participants and write on full sheets of paper the labels that you think are correct for your latent variables. Find people with some knowledge about the area and have them place the strips of paper on top of the label that they believe is most related to the question. Further, have them place a + on the strip if high scores on the latent variable should correspond to high scores on the questions, and - otherwise. You can then compare where participants placed the slips of paper with the loadings the computer produced (sometimes the latent variable is the opposite way around, so change the signs on all the paper strips for this exercise). You should get strips placed onto the labels which they load most highly on, and the direction of the loadings should make sense. If not, you should create different labels.

When reading journal articles that use latent variable techniques, you should consider whether the authors have used the same labels that you would have. Naming latent variables is a stage in the research process where researchers' biases are sometimes apparent.

Calculating Latent Variables

There are two reasons why people conduct latent variable analysis. The first is to understand the relationships among observed variables. The second reason is to create the latent variables, often called factor scores, and use these in other statistical procedures. There are several ways to calculate the factor scores. The most common method is the regression method. This will be shown in the second example in the next section. Do not be surprised if the latent variables have small correlations even if you use an orthogonal rotation like varimax. The correlations should be small, but they are unlikely to be exactly zero. If this bothers you, there is a method available in many packages called the Anderson-Rubin method which creates latent variables that have zero correlations. While this may seem an advantage, the simpler regression method is usually preferred (Tabachnick & Fidell, 2007). The correlations using the regression method are small enough that there will not be any problems with collinearity as described in Chapter *.

Part 2. Computing Examples and Output

Tabachnick and Fidell (2007) show how to conduct EFA and to interpret the output of EFA for SAS and SPSS, and also discuss SYSTAT. Field (2009) goes into more detail for SPSS users. These are both excellent sources. Here, we will show how to conduct and interpret EFA

using the freeware R and the package SPSS. Using freeware has the advantage that anyone can use it (details for downloading R are on our web site, but you can also google R), but because of the popularity of SPSS it is important to also understand how that program works too.³ We discuss a third program, CEFA (Browne et al., 2009), on the web page. CEFA is a free program that specializes in factor analysis. Thus, it can do more EFA than either R or SPSS, but it can be a hassle having to use a different package than for your main analysis. One advantage it has is that it produces standard errors for the factor loadings. Thus, if an instructor or editor asks for these, CEFA is a worthwhile program to learn.

We will examine 2 data sets. The first is one of datasets printed in Spearman's original paper. We use it to illustrate factor analysis with a single variable. The second uses some of our own data on social anxiety to show how to approach factor analysis when you do not know how many factors are appropriate.

Finding *g* (general intelligence): Small Sample with R

R can do a large number of different procedures and is extremely flexible. The R procedure for exploratory factor analysis, called **factanal**, has fairly standard output. We will show R commands and output in **bold Courier font**. There are fewer options with this function than the EFA procedures in the other packages including SPSS.

Table 2 shows some data from Spearman's original 1904 paper. He had 3 measures of people's ability to make discriminations using psychophysics methods (pitch, light, and weight) and four measures of attainment on school subjects (classics, French, English, and mathematics). Spearman felt that the observed variables were all related to a general intelligence factor. There were not appropriate methods then for this model, so he created factor analysis and performed the analysis. The sample size ($n = 23$) is much smaller than is now advised for EFA, but it is used here for illustration. Table 4 shows the correlations among these measures.

Insert Tables 2 and 3 about here

Since Spearman collected these data there have been many modifications to the basic factor analysis and its computing ease. Factor analysis can be easily run with many statistics packages. Computing packages also make it relatively easy to screen for rogue data and perform exploratory data analytic procedures. Before running an EFA you should examine histograms of the individual variables and look at the scatter plots between all pairs of variables. If there are some correlations among the variables then you can run a factor analysis. We go through some of these exploratory procedures on the chapter's web page, but not here due to space limitations.

Many packages allow factor analysis and most have similar output. The exact algorithms differ so the precise numeric output may differ, but the overall conclusions should be the same. We provide more details of all analyses on the web page. The data are stored in an object called **Speardata**. In R, to run a factor analysis looking for one factor you write:
factanal(Speardata, factors=1)

This produces a lot of output which we will go through in sections. The first section of the output is the uniqueness scores. These are one minus the communalities. The total variation for each

³ SPSS was bought by IBM and is going through some changes, including a new name: PASW (for predictive analytic software), though it is still called SPSS by many cool people.

observed variable can be divided into the proportion accounted for by the shared factors (the communality) and the proportion not shared (the uniqueness).

Each of the original variables' variation can be divided into that unique to the variable and the shared with the latent variable(s). R prints the unique variance, called uniqueness. The high values for pitch (0.939), light (0.993), and weight (1.000) show that almost all the variation of these variables is unique to each of these variables. Very little is shared with the latent variable. The amount shared with the latent variable is the communality. These values are near zero for these three variables. The variables classics, French, English, and math all have low uniqueness values and high communalities, showing they are related to the latent variable. In sum, the uniqueness scores show that the single factor does not account for psychophysical discrimination measures (pitch, light, and weight), but does account for the academic measures.

Uniquenesses:

pitch	light	weight	classics	french	english	math
0.939	0.993	1.000	0.044	0.070	0.350	0.220

To find the communalities in R, type:

```
1-factanal(Speardata,1)$uniqueness
```

and you get (we have printed only a few of the digits):

pitch	light	weight	classics	french	english	math
0.061	0.007	0.000	0.956	0.930	0.650	0.780

or just subtract each uniqueness score from 1 by hand.

The loadings show that the factor loads highly on all the academic measures, and particularly on classics and French. The blank loadings, for light and weight discrimination, are less than .2 and the computer by default does not print these. The loading for pitch is also very low.

Loadings:

	Factor1
pitch	0.247
light	
weight	
classics	0.978
french	0.965
english	0.806
math	0.883

With a one-factor model the square of the loadings is the communality. Thus, $.247^2 = .061$, which is the communality for pitch.

The proportion of variance accounted for by the single factor is 48.3% or 3.383 eigenvalues ($3.383/7 = .483$).

	Factor1
SS loadings	3.383
Proportion Var	0.483

Test of the hypothesis that 1 factor is sufficient.

The chi square statistic is 13.01 on 14 degrees of freedom.

The p-value is 0.526

This final part of the output shows that the model fits, $\chi^2(14) = 13.01$, $p = .526$. If this were a larger sample it would suggest that a single factor model could be appropriate.

To create the factor scores, using the regression method, type

```
factanal(Speardata,1,scores="regression")$scores
```

which will produce the estimated g for each person. If you want to calculate the correlation matrix implied by this single latent variable model type:

```
alpha <- factanal(Speardata,1,scores="regression")$loadings
```

to create the loadings. Then to create the implied correlation matrix type:

```
impCORR <- alpha %*% t(alpha)
```

To make this so it has ones on the diagonal, type:

```
diag(impCORR) <- 1
```

and then type:

```
impCORR
```

to show the implied correlation matrix. To see how much this is off from the observed correlation matrix type: `cor(Speardata) - impCORR`. We do not print these matrices here, but they are on our web page. The model suggests low correlations for weight with pitch ($r = -.003$) and weight with light ($r = -.001$). The observed correlations were $r = .203$ and $r = .314$, respectively, so the model is poor at recreating these.

If you only want the loadings that are above some value in magnitude, say .20, then in the above code write:

```
alpha[abs(alpha) < .2] <- 0
```

after `alpha` is defined and repeat the steps from above. The two-factor model is found with

```
factanal(Speardata,factors=2) and the output is:
```

Uniquenesses:

pitch	light	weight	classics	french	english	math
0.936	0.847	0.615	0.071	0.005	0.271	0.181

Loadings:

	Factor1	Factor2
pitch	0.252	
light		0.384
weight		0.619
classics	0.961	
french	0.992	-0.108
english	0.791	0.321
math	0.885	0.188

	Factor1	Factor2
SS loadings	3.387	0.687
Proportion Var	0.484	0.098
Cumulative Var	0.484	0.582

Test of the hypothesis that 2 factors are sufficient.
 The chi square statistic is 4.76 on 8 degrees of freedom.
 The p-value is 0.783

The χ^2 for the one-latent variable model was 13.01 with 14 degrees of freedom. The χ^2 for the two-latent variable model is 4.76 with 8 degrees of freedom. The difference in degrees of freedom (6) is the number of observed variables minus 1, here 7-1. The difference between the χ^2 from this output (4.76) and that found in the output for the single latent variable model (13.01) values is 8.25. With $df=6$ the p value is .22. Thus, adding the extra latent variable does not significantly improve the fit of the model, but with the small n this is expected.

Also, the proportion of variation accounted for by the second latent variable is much less than the first (10% versus 48%) and is less than 1/7 (which is 14% and for this example is one eigenvalue). Thus, this supports (albeit with a small sample) Spearman's view of a single latent variable. To create factor scores for the 2 latent variable model type:

```
factanal(Speardata,2)$scores
```

If a second or third latent variable was needed, then the loadings could have been rotated. For varimax rotation, in R, write: **varimax(factanal(Speardata,2)\$loadings)** Because the single latent variable model fits better than the two latent variables model, you would not need to use rotation here. The web page shows how to conduct these analyses with SPSS and CEFA, and compares the output of the three packages.

How large a sample is necessary for EFA?

As with many statistical rules of thumb the answer is complex and it depends on many things including why you are doing the EFA. Field (2009) reviewed several sources and found that samples above 300 are usually safe. With smaller samples, if a factor has several loadings above .6 then that factor probably has some value. With the Spearman g in the example loading highly on 4 variables this suggests that there is some common element to those 4 variables.

Social Anxiety: Multiple Factor Analysis in SPSS

We research how easy it is (and sometimes is not) to change people's memories. It turns out that social anxiety may relate to this. We took 11 questions from a large set of items and will use these to illustrate the two-factor model for 68 participants from Villalba et al. (under review). The variables were from two scales, one purporting to measure fear of negative evaluation (items we label BFNE) and one purporting to measure social avoidance (SIAS). Therefore we thought that there might be different latent variables which influence these items, but we were not sure. We will go step by step through this example using SPSS. The data are available on the chapter's web page along with analyses in R and CEFA. The items are labeled: BFNE_01, BFNE_05, BFNE_06, BFNE_08, and BFNE_11; and SIAS_11, SIAS_13, SIAS_14, SIAS_15, SIAS_17, and SIAS_18.

The first step would be to look at the histograms of the individual variables and scatter plots between variables (Chapter *). Assuming all this is done and there are no strange outliers, the next step is to see if there is a correlation between the observed variables. As we have stated at the beginning of the chapter, if the observed variables are not correlated with each other then there is no need to do an EFA. Within SPSS the correlation matrix can be printed from the procedure that runs EFA or from the correlation procedure. We advise using the correlation procedure to ensure that you look at the correlations prior to conducting EFA. Details about the correlation statistic are in Chapter *. Within SPSS go to Analyze, then Correlate, then Bivariate (the exact words vary among versions). Then you drag all the variables you want to correlate

from the left column to the right one (labeled Variables). The default is Pearson's correlation. The default is to print *s for "significant" correlations. Given the number of correlations which are all related here (because the variables are all correlated) it is unclear what any of these mean so you may want to untick the box that asks you if you want stars by the "significant" correlations. Next, press OK. The output will look like Figure 6. We have changed the font to make it more easily read on a book's page and also drawn on it ourselves. With correlation matrices it is worth printing them off and looking at them! We started by drawing a line through the diagonal, because all these are 1. The correlations are the same above and below the diagonal, so we drew a squiggly line to show to ignore the ones below the diagonal. We drew lines to block off those correlations within the BFNE variables and with the SIAS variables, and a box around the correlations between these sets. Because all the questions are about social anxiety, they are all positively correlated, but it looks like the correlations within the same set are higher. We are ready to run EFA.

Insert Figure 6 about here

To conduct an EFA in SPSS go to Analyze then Dimension (or Data) Reduction and then Factor. Again, the exact phrases change between SPSS/PASW versions. This will open a dialog box like Figure 7. Place the variables which you want to include within the EFA into the box labels Variables. The tabs on the upper right side (Descriptives, Extraction, Rotation, Scores, Options) of Figure 7 allow you to change different aspects of how the EFA is conducted and what output is reported. Field (2009) goes through in detail on the options. Here we discuss just some of the options available.

Insert Figure 7 about here

The Descriptives tab allows you to print each variable's mean and standard deviation, and the correlation matrix, all of which you should have already examined. It also allows you to print the implied correlation matrix (labeled Reproduced). This is a useful option so we ticked it. The Extraction tab has many useful facets. First, you need to decide the method to use. We use Maximum likelihood so that our results are comparable with other packages. The SPSS default, at least in current versions, is to use principal components analysis, which most statisticians will be quick to point out is actually different from factor analysis. The other methods are all techniques that statisticians have argued can be useful in some circumstances. This tab also allows you to print a scree plot, which we always recommend. It also allows you to say how many factors you want. Usually you would decide this after you look at the scree plot and other output (which means you often run a series of EFAs). Rotation allows you to try different rotations. We ticked Varimax. The Scores tab allows you save the factor scores. Regression method is the most popular. This would allow these scores to be used in other analyses. The Options tab allows missing values to be filled in (though we recommend using procedures specifically designed for this) and to suppress small loadings. For this example we do not use these.

The complete output is shown on the web page but we will discuss some of the important parts here. The first information output is the communalities:

Communalities

	Initial	Extraction
BFNE_01	.583	.533
BFNE_05	.697	.696
BFNE_06	.533	.519
BFNE_08	.744	.793
BFNE_11	.644	.608
SIAS_11	.565	.529
SIAS_13	.613	.503
SIAS_14	.677	.673
SIAS_15	.706	.808
SIAS_17	.688	.591
SIAS_18	.723	.770

Communalities: The proportion of variance of each observed variable that is accounted for by the factors. Thus, for BFNE_01 the two factor solution (which is what is reported later in the output) accounts for 53.3% of the variation. A variable that shares none of its variance with any other variable has a communality of zero. If a communality is near 0 the variable can be removed. If a communality is near or above 1 it is a sign that there may be computational difficulties.

Extraction Method: Maximum Likelihood.

Next SPSS outputs the variance accounted for by each factor. The output is too wide to print legibly on page, but is included on the web. The information in the table is used to create the scree plot which is shown in Figure 8 (after using the chart editor). The scree suggests 1 or 2 factors. Both of the first two factors have eigenvalues above 1 (6.225 and 1.453, respectively), while the eigenvalue associated with the third factor is below 1 (.712). Because we have theoretical reasons to believe the questions are from two related but different aspects of social anxiety we would probably choose the 2 factor solution.

Insert Figure 8 about here

SPSS prints a χ^2 value for this two-factor solution: $\chi^2(34) = 54.534$. This is most useful in comparison with other χ^2 values. The solution for the one-factor model is: $\chi^2(44) = 146.261$. The reduction is statistically significant, change in $\chi^2(10) = 91.727, p < .001$.

Next the unrotated factor loading matrix is produced. Here we show it with the rotated matrix (though the rotated matrix is printed near the end of the output, it is useful to present here for comparison). For the unrotated solution all the variables load on the first factor. The second factor has positive loadings for the fear of negative evaluation questions and mostly negative loadings for the social avoidance questions. For the rotated solution the fear of negative evaluation questions load mostly on the first factor and the social avoidance questions mostly on the second. The overall fit of the model does not change with rotation nor do any of the communalities (e.g., $(.634)^2 + (.361)^2 = (.696)^2 + (.219)^2 = .53$).

Factor Matrix(a)

	Factor	
	1	2
BFNE_01	.634	.361
BFNE_05	.763	.337
BFNE_06	.682	.232
BFNE_08	.714	.532
BFNE_11	.716	.309
SIAS_11	.720	.103
SIAS_13	.706	-.073
SIAS_14	.784	-.241
SIAS_15	.743	-.506
SIAS_17	.739	-.209
SIAS_18	.806	-.346

Extraction Method: Maximum Likelihood.
a 2 factors extracted. 5 iterations required.

Rotated Factor Matrix(a)

	Factor	
	1	2
BFNE_01	.696	.219
BFNE_05	.767	.330
BFNE_06	.634	.342
BFNE_08	.876	.161
BFNE_11	.714	.314
SIAS_11	.565	.458
SIAS_13	.426	.567
SIAS_14	.357	.738
SIAS_15	.134	.889
SIAS_17	.350	.684
SIAS_18	.295	.826

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.
a Rotation converged in 3 iterations.

SPSS prints the transformation matrix:

Factor Transformation Matrix

Factor	1	2
1	.680	.733
2	.733	-.680

Extraction Method: Maximum Likelihood.
Rotation Method: Varimax with Kaiser Normalization.

Multiplying the unrotated matrix by this rotation matrix produces the rotated factor matrix shown above (so $(.634 \cdot .680) + (.361 \cdot .733) = .696$). The data are rotated: $\arccos(.68) = \arcsin(.73) = .818$ in radians, which if we multiply by $180/\pi$ we get a shift of 47° . The rotated loadings can be plotted either within SPSS or from the values above, see Figure 9. As can be seen the variables all still load positively on the factors. We included arrows from the origin to the mean factor loadings for the SIAS and BFNE variables. Because these arrows are not at right angles to each other it means the scores on these sets of items are correlated. This would be a situation where a non-orthogonal rotation might be used. Non-orthogonal rotations allow the factors to be correlated. As said above, these are mathematically more complex. However computationally it is relatively simple because the computer does it. You choose one of the non-orthogonal methods (in SPSS Oblimin or Promax) and how much correlation to allow (try a few values). This is explored more on the web page.

Insert Figure 9 about here

Finally, SPSS prints the correlation matrix implied by the model (shown on the web page). Here it is the two-factor model. The implied matrix is the same whether the data are

rotated or not. SPSS also prints the difference between the implied matrix and the sample matrix. Only 2 of the 45 correlations were off by more than .1, which shows that the model captures most of the associations among variables fairly well. The two slightly errant ones were the model assumes BFNE_01 and BFNE_11 are correlated .57 (sample value .69) and that SIAS_13 and SIAS_17 are correlated .59 (sample value .38). The three-factor model also has two correlations also off by more than .1.

In summary, this example shows running an EFA in SPSS when it is expected that a two-factor model may fit. In this sense the analysis is not purely exploratory, but there are very few cases where researchers have no a priori beliefs about how many factors there are. Here we examine whether a two factor model fits better than other models using the scree plot, significance testing (the χ^2 tests), and how well the model reproduced the correlation matrix. Further, we looked at the rotated loadings to see if the EFA split the variables in a similar manner to our a priori beliefs. They did (although one of the SAIS items is close to the mean of the BFNE items so this item could be re-classified or discarded).

3. Extensions to Exploratory Factor Analysis

We have talked about one type of latent variable approach: exploratory factor analysis or EFA. It is appropriate when you hypothesize that a small number of interval latent variables are responsible for variation in a larger number of observed variables, but you are unsure how they will be related. Within psychology EFA is the most used type of latent variable model. Here we briefly discuss some extensions to this model. We begin with models where you have strong *a priori* beliefs about the relationships among latent and observed variables and you want to either confirm or disconfirm your beliefs about the particular relationships.⁴ Next, we describe models appropriate when either the observed or the latent variables are categorical.

Confirmatory Methods

An alternative to exploratory factor analysis is confirmatory factor analysis, abbreviated CFA. In EFA the researcher does not specify how the latent variables are related to the observed variables. In CFA the researcher specifies these relationships. They use the diagrams with ellipses for latent variables and rectangles for observed variables, and *they* add in the arrows, and then the computer estimates the loadings for those arrows.

EFA: Computer decides where to put the arrows.

CFA: Researcher decides where to put the arrows.

CSI: Detectives figure out who shot the arrows.

The computer also provides a measure of how good the fit is. The researcher usually runs several models and decides which models are inconsistent with the data and which ones fit well. The purpose is usually to discard bad models than to decide among a set of good models.

With CFA researchers often hypothesize arrows among latent variables to show the variables are either correlated or that one is influencing another. These models are sometimes called structural equation models, or SEMs. They are mathematically more complex than EFA, and also require more care when conducting them with a statistics package. See Klem (2000), Miles and Shevlin (2003), and Thompson (2000) for introductions. The analyst has to make so many assumptions with some of these models that some methodologists argue that SEM models should be used sparingly if at all. Bartholomew and Knott (1999, p. 190) end their seminal work

⁴ As noted in the last example researchers always have beliefs about the relationship among variables. In some sense it is a matter of the degree of belief and in some cases personal preference for determining whether exploratory or confirmatory methods are used.

on latent variables with: "When we come to models for relationships between latent variables we have reached a point where so much has to be assumed that one might justly conclude that the limits of scientific usefulness have been reached if not exceeded".

Levels of Measurement of Latent and Observed Variables

In Chapter * Fife-Shaw described the levels of measurement framework to differentiate types of variables. Both the latent and observed variables could be any of these levels. The most commonly assumed levels are interval and categorical. Table 4 shows Bartholomew and colleagues (2009) classification of different latent variable models for these different levels of variables. A different name is given to each method. The methods were developed relatively independently of each other. Bartholomew and Knott (1999) show how each is a special case of what they call the *general linear latent variable model*. Factor analysis was mainly developed in psychology and assumes that both the latent and observed variables are interval. The latent class and latent profile methods were developed mainly in sociology. Latent class analysis assumes categorical latent and observed variables. Latent profile analysis assumes interval observed variables and categorical latent variables. Latent profile analysis is rare in psychology, where other techniques (e.g., taxometric analysis, discriminant function analysis, cluster analysis) are used in these situations. Latent class models are the more common than latent profile analysis, particularly in sociology.

Insert Table 4 about here

The latent trait models, also called item response models, are common in education. The most common use of latent trait models are where you have multiple items where people either pass or fail on each item and you assume that there is one underlying latent variable (called a trait for these models rather than a factor). In fact, when students at our university take a multiple choice exam, the feedback given to the instructor includes output from this latent trait model. The instructor could use scores on this latent trait as a measure of achievement. The output can also help them to decide which questions are good and which are bad. Instructors can omit responses to bad questions from people's grades for that year and hopefully remove the offending questions from subsequent exams. Embretson and Reise (2000) have written an excellent text on these models.

Trying to decide whether your latent variables are categorical or metrical should be based largely on your theories rather than on the data. This is because often these approaches will fit both types equally well (Bartholomew, 1993). In general finding that a statistical model fits the data does not mean it is a good model. It is important to see how it fits compared with others. Meehl and colleagues (e.g., Waller & Meehl, 1998) have developed a set of methods to help decide whether data are more consistent with metrical or categorical latent variables, but these methods require a fairly large amount of data. Links to various sources are on the web page.

Principal Components Analysis

A common alternative to EFA is principal components analysis, or PCA. PCA works by reducing a large number of observed variables into a smaller number of components where the components account for a large amount of the variation of the original variables. Thus, its purpose is very similar to EFA. PCA is a data reduction technique that does not make the assumption that latent variables exist. The equations for PCA are almost the opposite as those for EFA: the components are a linear combination of the observed variables (in PCA) rather than the observed variables being a linear combination of the factors (in EFA). However, both procedures work by accounting for variation among the observed variables. The solutions of the two are so similar that some statistical packages, including SPSS, simply list PCA as one way to estimate a factor analysis.

Often the two procedures are combined. A PCA is used to create a scree plot to help decide how many latent variables are needed (for technical reasons making a scree plot with PCA is more straightforward), and then an EFA is applied.

4. Summary

There seems to be something sneaky about exploratory factor analysis. How can a data analysis technique somehow allow you to measure what has not been measured? There is a South Park episode where Butters believes he has become a vampire. In part this was to allude being grounded because he believes vampires cannot be grounded. He poses the deep philosophical question to his parents: *"How can thy ground that which is ungroundable?"* Similarly, how can we calculate factor scores and do any analyses with latent variables if they are un-measurable? This conundrum has meant many statisticians have been skeptical of factor analysis. Butters solved his problem by helping a group of Goths burn down a clothing shop. The way we circumvent the latent variable problem is because the question we are asking is after an assumption. First, we assume that the latent variables exist, and then we ask the computer: *"If they exist, what loadings are most consistent with the data?"* There still is a bit of hand waving, particularly when naming the latent variables, but the approach is now generally accepted within statistics. Further, it has been helpful for a hundred years of psychology. This does not mean it is right, but it does mean that it is a tool worth keeping in your statistical arsenal.

Most psychologists do not dwell on philosophical problems with statistical approaches, but ask what the approach can do for them. EFA can help a) to reveal exciting patterns in their data, and b) to tell a good story about their data. The EFA solution shows how the different observed variables hang together. It allows the researcher to summarize their data by assuming the latent variables map onto relevant psychological constructs. Many research papers use EFA of questionnaires show patterns in the responses. The second main use of EFA is to create the latent variables or factor scores, and to use these in subsequent analyses. If an orthogonal rotation, like varimax, is used then these subsequent factors are essentially uncorrelated. This helps to provide a simple solution and avoids the collinearity problems discussed in Chapter *.

Exploratory Factor Analysis has passed its hundredth birthday and is as strong as ever. The initial years were difficult because the technique had to be conducted without computers. In the past few decades advances in computing technology has lead to the widespread use of many complex multivariate techniques including factor analysis. A lot of this chapter is based on the

first author's student-days listening to David Bartholomew's brilliance. Two things he said are most important here. First, he started his multivariate statistics course saying his goal was to stop us using complex multivariate procedures. He said how computing (even back then) made it too easy first to run your analyses and only then to think about what analyses you should have run. With a procedure as complex as EFA it is critical to think about what statistics you will do before doing them. His second dictum was that with many multivariate procedures the proof is in the eating. If the result of the analyses shows an illuminating pattern in the data, then that is good. For EFA this means trying a few estimation methods, a few rotations, look at different numbers of latent variables, and find which of the many solutions is most revealing.

5. Exercises

1. A sample of 100 people was asked to respond to the 8 questions. Suppose the resulting loadings for the unrotated model were:

Loadings:

	Factor1	Factor2
[1,]	0.503	0.483
[2,]	-0.524	0.652
[3,]	0.760	0.205
[4,]	0.774	0.213
[5,]	-0.314	0.202
[6,]	0.603	0.415
[7,]	-0.369	0.402
[8,]	-0.504	0.473

The rotated loadings (using varimax) are:

Loadings:

	Factor1	Factor2
[1,]	0.693	0.074
[2,]	-0.015	0.837
[3,]	0.727	-0.303
[4,]	0.742	-0.305
[5,]	-0.125	0.352
[6,]	0.731	-0.041
[7,]	-0.045	0.544
[8,]	-0.110	0.683

Create diagram for both of these, using ellipses, rectangles, and arrows, to show the relationships between the latent factors and the observed variables. Which diagram appears simpler? The answer may be affected by how large a loading you decide you need for drawing an arrow.

2. The rotation matrix for exercise #1 was: $\begin{bmatrix} .79 & -.61 \\ .61 & .79 \end{bmatrix}$. How much were the data rotated?

Describe why you would do rotation.

3. Describe the methods that you use to decide the number of factors.
4. . What makes people happy? A recent movement called positive psychology examines what makes people happy. Suppose a survey researcher took a random sample of 1000 people and asked them 10 attitudinal questions (called att1 to att10) and gave them a happiness questionnaire (happy). The data are saved in various formats on our web page.
 - a. Describe the relationships among the different attitudinal measures using correlation and EFA.
 - b. If you have already done multiple regression, save the factor scores from an EFA, and use these to predict happiness.

6. Questions

5. EFA is often used to simplify how the relationships among many variables are described. Discuss why simplifying complex patterns can be helpful in science. Describe what also must be considered so that the account is not too simplified. State one interesting biographical detail about William of Ockham.
6. Suppose a researcher believed that some people are simply happy (see <http://www.sohp.com/>), and some are not. Could EFA be used to help the research empirically validate this belief? If so, how?

7. Suggested Readings

These readings are roughly in order of assumed mathematical experience (Tabachnick and Bartholomew are close).

Chapters with emphasis on computing packages

1. Field, A.P. (2009). *Discovering statistics using SPSS (and sex and drugs and rock'n'roll)* (3rd ed.). London: Sage Publications. Great descriptions of all the SPSS options available for EFA (the statistics are serious, but some of jokes you might not share with your parents).
2. Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon. A further survey of most of the statistical techniques psychologists learn in the first year of graduate school.

Book with a chapter on EFA but also on many of the alternatives

3. Bartholomew, D.J., Steele, F., Moustaki, I., & Galbraith, J.I. (2008). *Analysis of multivariate social science data* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC. This textbook, aimed at first year social science graduate students, provides the best overview of the different types of latent variable models in a manner accessible to non-statisticians. Bartholomew and Knott (1999) provides coverage aimed at applied statisticians.

Comprehensive book the mathematics of factor analysis

4. Mulaik, S.A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC. This is a thorough and detailed review of factor analysis, included details of many of the controversial issues that are skirted over in introductory texts and single chapters. This book requires more mathematical knowledge than the other texts.

Glossary

Communalities are the proportion of variability in the observed variables accounted for by the factors. You can think of them like a multiple R^2 predicting the observed variable from all the latent variables.

An **eigenvalue** is a mathematical term used in calculating several advanced statistics procedures. For our purposes it is a unit of measurement for how much variation is accounted for by the latent variables.

Factor scores are the values for each participant for the hypothesized factors. These can be created with all EFA software.

Implied correlation matrix is the matrix that the model is able to recreate. This should be close to the observed correlation matrix, otherwise the model does not fit the data well.

Item specific errors are the errors assumed to apply separately to each variable. They are the *es* in circles added to some path/causal diagrams.

Latent variables are those we do not directly observe. The type of latent variable discussed in this chapter is usually called a factor.

Loadings, denoted with the Greek letter α or the letter L, show the strength between the latent variable and the observed variable.

Manifest or observed variables are those which are directly measured.

Rotation is used to make the loadings easier to interpret. *Varimax* is the most common orthogonal rotation.

The **scree plot** is a line graph of additional variation accounted for with the number of latent variables. It is useful for decide how many latent variables are needed. The scree plot is often made using a procedure called principal components analysis, which is described at the end of this chapter.

References

- Abelson, R. P. (1995). *Statistics as principled argument*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Bartholomew, D.J., Steele, F., Moustaki, I., & Galbraith, J.I. (2008). *Analysis of multivariate social science data* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Bartholomew, D. J. & Knott, M. (1999). *Latent variable models and factor analysis (Kendall's Library of Statistics 7)*. London: Arnold.
- Browne, M.W., Cudeck, R., Tateneni, K., & Mels, G. (2009). *CEFA: Comprehensive Exploratory Factor Analysis* (version 3.03). Available <http://faculty.psy.ohio-state.edu/browne/software.php> (Sept, 2010).
- Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

- Field, A.P. (2009). *Discovering statistics using SPSS (and sex and drugs and rock'n'roll)* (3rd ed.). London: Sage Publications.
- Klem, L. (2000). Structural equation modeling. In L G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 227-260). Washington, DC: American Psychological Association.
- Miles, J. & Shevlin, M. (2003). Structural equation modelling: Navigating spaghetti junction. *The Psychologist*, **16**, 639-641.
- Mulaik, S.A. (2010). *Foundations of factor analysis* (2nd ed.). Boca Raton, FL: Chapman & Hall/CRC.
- Nova Development Corporation. (2004). *Art explosion: 800,000 premium quality graphics* [CD Rom].
- Spearman, C. (1904). "General intelligence," objectively determined and measured. *American Journal of Psychology*, *15*, 201-293. Available <http://psychclassics.asu.edu/Spearman/> (Sept. 2010).
- Tabachnick, B.G., & Fidell, L.S. (2007). *Using multivariate statistics* (5th ed.). Boston: Allyn & Bacon.
- Thompson, B. (2000). Ten commandments of structural equation modeling. In L G. Grimm & P. R. Yarnold (Eds.), *Reading and understanding more multivariate statistics* (pp. 261-283). Washington, DC: American Psychological Association.
- Thompson, B. (2004). *Exploratory and confirmatory factor analysis: Understanding concepts and applications*. Washington, DC: American Psychological Association.
- Villalba, D.K. & Wright, D.B. (under review). Informational influences on co-witness memory.
- Waller, N. G. & Meehl, P. E. (1998). *Multivariate taxometric procedures: Distinguishing types from continua*. Thousand Oaks, CA.: Sage Publications.

The first entry in the product ($ag + bk$) is found by multiplying the first value in the first row of the first matrix (a) by the first value in the first column of the second matrix (g), and then repeating this with all other values in this row and this column (here only 2), and adding these up. For this to work there has to be the same number of values in each row of the first matrix (i.e., the number of columns) as the number of values in each column of the second matrix (i.e., the number of rows). The resulting matrix has the same number of rows as the first matrix and the same number of columns as the second matrix.

One way to remember this is that in order to multiply two matrices, if you place their dimensions side-by-side, like above with the 3x2 and 2x4, then the inner dimension values must be the same (the 2s) and the outer dimension values become the dimensions for their product. In our example the \mathbf{F} matrix is 10x2 so we can only multiply that by a matrix with 2 rows. The $\boldsymbol{\alpha}$ matrix is 8x2, so we cannot estimate the product $\mathbf{F}\boldsymbol{\alpha}$. However, we can transpose the $\boldsymbol{\alpha}$ matrix and the transposed matrix is 2x8. Once you transpose the matrix you can multiply it.

$$\boldsymbol{\alpha}^T = \begin{bmatrix} \alpha_{11} & \alpha_{21} & \alpha_{31} & \alpha_{41} & \alpha_{51} & \alpha_{61} & \alpha_{71} & \alpha_{81} \\ \alpha_{12} & \alpha_{22} & \alpha_{32} & \alpha_{42} & \alpha_{52} & \alpha_{62} & \alpha_{72} & \alpha_{82} \end{bmatrix}$$

The model can be written out as: $\mathbf{X} = \mathbf{F}\boldsymbol{\alpha}^T + \mathbf{E}$. This is much shorter than trying to write out the large number of equations without using matrix notation.

In EFA you often want to calculate the correlation matrix implied by the loadings. This is:

$$\boldsymbol{\alpha}\boldsymbol{\alpha}^T = \begin{bmatrix} \alpha_{11} \\ \alpha_{21} \\ \alpha_{31} \\ \alpha_{41} \\ \alpha_{51} \\ \alpha_{61} \\ \alpha_{71} \\ \alpha_{81} \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{21} & \alpha_{31} & \alpha_{41} & \alpha_{51} & \alpha_{61} & \alpha_{71} & \alpha_{81} \\ \alpha_{12} & \alpha_{22} & \alpha_{32} & \alpha_{42} & \alpha_{52} & \alpha_{62} & \alpha_{72} & \alpha_{82} \end{bmatrix}$$

$$= \begin{bmatrix} \alpha_{11}\alpha_{11} & \alpha_{11}\alpha_{21} & \alpha_{11}\alpha_{31} & \alpha_{11}\alpha_{41} & \alpha_{11}\alpha_{51} & 0 & 0 & 0 \\ \alpha_{21}\alpha_{11} & \alpha_{21}\alpha_{21} & \alpha_{21}\alpha_{31} & \alpha_{21}\alpha_{41} & \alpha_{21}\alpha_{51} & 0 & 0 & 0 \\ \alpha_{31}\alpha_{11} & \alpha_{31}\alpha_{21} & \alpha_{31}\alpha_{31} & \alpha_{31}\alpha_{41} & \alpha_{31}\alpha_{51} & 0 & 0 & 0 \\ \alpha_{41}\alpha_{11} & \alpha_{41}\alpha_{21} & \alpha_{41}\alpha_{31} & \alpha_{41}\alpha_{41} + \alpha_{42}\alpha_{42} & \alpha_{41}\alpha_{51} + \alpha_{42}\alpha_{52} & \alpha_{42}\alpha_{62} & \alpha_{42}\alpha_{72} & \alpha_{42}\alpha_{82} \\ \alpha_{51}\alpha_{11} & \alpha_{51}\alpha_{21} & \alpha_{51}\alpha_{31} & \alpha_{51}\alpha_{41} + \alpha_{52}\alpha_{42} & \alpha_{51}\alpha_{51} + \alpha_{52}\alpha_{52} & \alpha_{52}\alpha_{62} & \alpha_{52}\alpha_{72} & \alpha_{52}\alpha_{82} \\ 0 & 0 & 0 & \alpha_{62}\alpha_{42} & \alpha_{62}\alpha_{52} & \alpha_{62}\alpha_{62} & \alpha_{62}\alpha_{72} & \alpha_{62}\alpha_{82} \\ 0 & 0 & 0 & \alpha_{72}\alpha_{42} & \alpha_{72}\alpha_{52} & \alpha_{72}\alpha_{62} & \alpha_{72}\alpha_{72} & \alpha_{72}\alpha_{82} \\ 0 & 0 & 0 & \alpha_{82}\alpha_{42} & \alpha_{82}\alpha_{52} & \alpha_{82}\alpha_{62} & \alpha_{82}\alpha_{72} & \alpha_{82}\alpha_{82} \end{bmatrix}$$

This looks really complicated, but if you are okay with matrix algebra it is just the product of two matrices. It is easily produced in the statistics packages as shown in the text and on the web page.

This has been a brief introduction to matrices. Most multivariate statistics books written for psychologists have a chapter on matrix computations. Sources are given on the web page.

Table 1. Sample questions that should be influenced by whether somebody is or is not socially avoidant.

If I could move anywhere, it would be:

1	2	3	4	5	6	7
Inside a South Beach Club					A tent in northern Greenland	

When I watch football, I like to:

1	2	3	4	5	6	7
Go early to the stadium and enjoy the crowd				Watch on my 63" screen with my own tortilla chips		

My idea of a perfect Saturday afternoon is:

1	2	3	4	5	6	7
Going to a massive rave and getting high				Sitting by a quiet lake and getting high		

The statement that best describes my use of a cell/mobile phone is:

1	2	3	4	5	6	7
I txt left-handed while driving					I have it so my mom can call	

Table 2. The data from Spearman (1904, Experimental Series 3a).

Discrimination thresholds			Place in School (ranks)			
Pitch	Light	Weight	Classics	French	English	Mathem.
50	10	4	16	19	10	7
3	10	6	5	6	6	5
10	10	9	13	11	11	13
>60	10	9	22	23	22	22
4	12	5	1	1	1	2
2	10	10	4	2	2	1
4	10	11	12	14	13	18
20	10	11	23	22	23	23
11	10	12	8	8	15	15
11	12	11	3	5	4	4
24	14	10	7	7	7	6
5	18	7	20	15	18	16
3	18	9	10	13	14	12
5	13	13	2	3	3	3
6	13	13	11	12	12	9
7	14	11	17	18	17	13
15	19	10	21	20	21	19
11	14	13	19	21	9	21
14	13	18	18	16	8	17
15	13	28	15	10	20	10
7	19	13	9	9	16	11
4	16	16	14	17	19	20
>60	19	27	6	4	5	8

The ">60" values in the pitch column are treated as the value 60.

Table 3. The correlation matrix for the data in Table 2. There are high correlations among classics, French, and English, and small correlations elsewhere.

	pitch	light	weight	classics	French	English
light	-.02					
weight	.20	.31				
classics	.26	.09	.02			
French	.26	.04	-.11	.94		
English	.11	.19	.10	.79	.75	
Math	.13	.11	.09	.85	.86	.78

Table 4. Bartholomew's classification of latent variable models.

		Observed variables	
		Interval	Categorical
Latent variables	Interval	Factor analysis	Item response models (Latent trait analysis)
	Categorical	Latent profile analysis	Latent class analysis

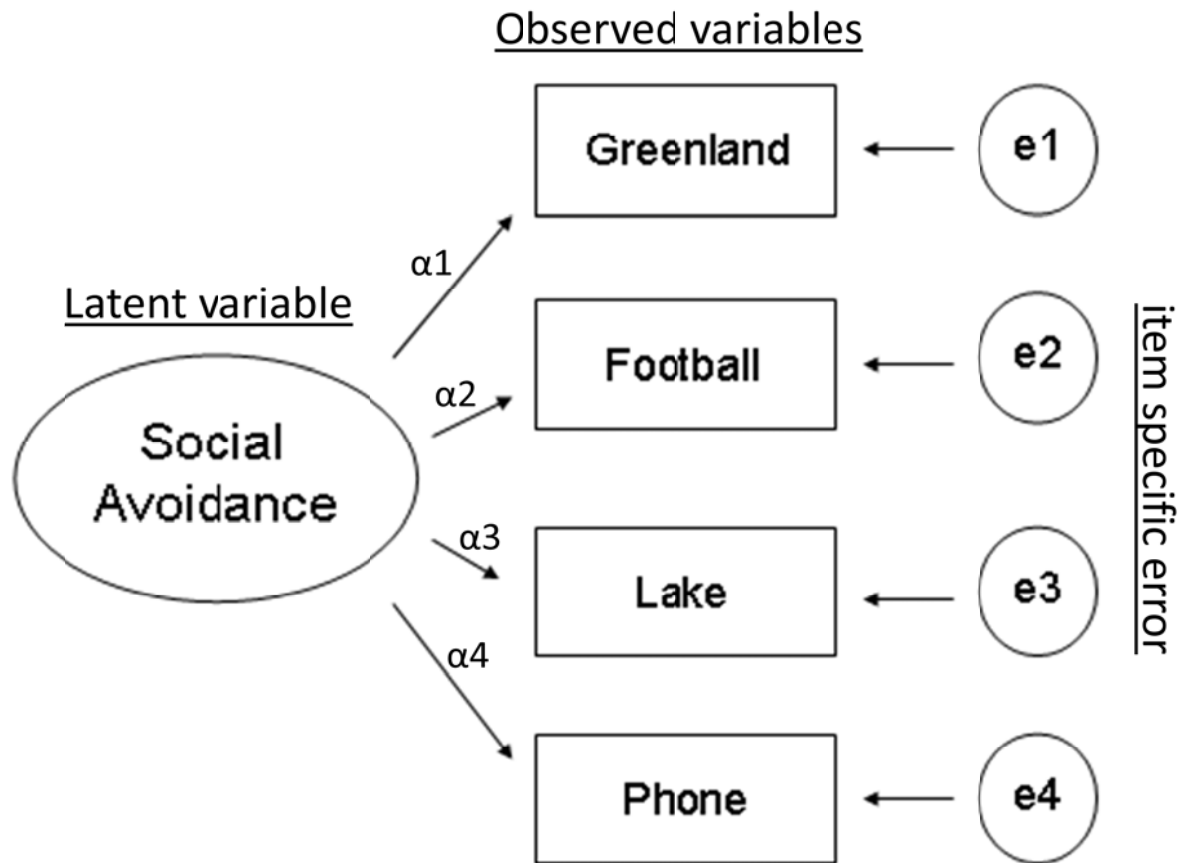


Figure 1. A causal diagram showing one latent variable affecting the responses for four observed variables, which are also each influenced by its own item specific error term.

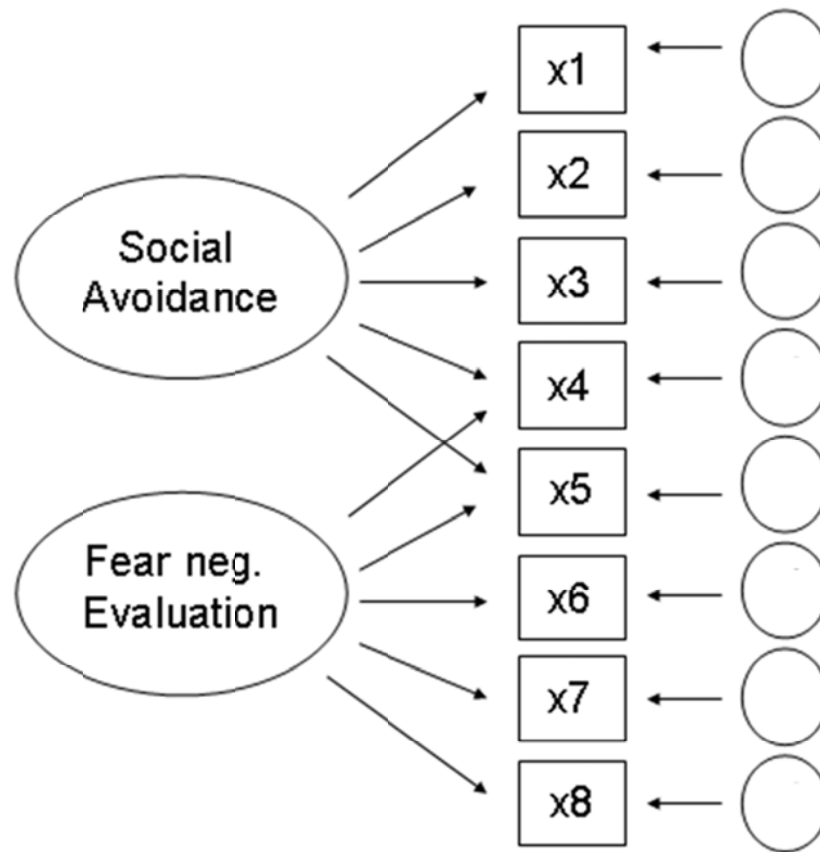


Figure 2. A two latent variable model which shows "Social Avoidance" affecting responses to observed variables x1 to x5, and "Fear of Negative Evaluation" affecting responses to x4 to x8.

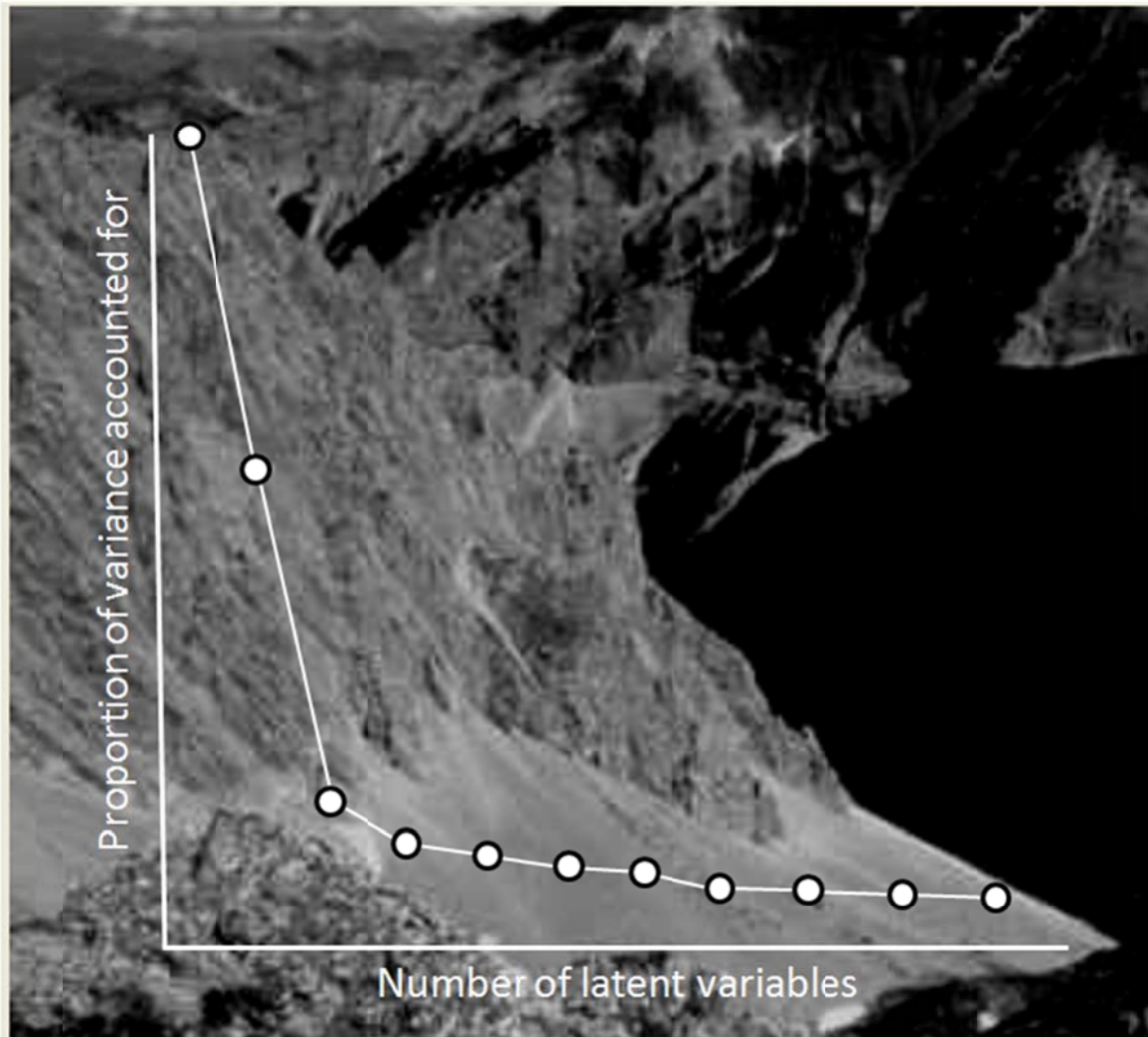
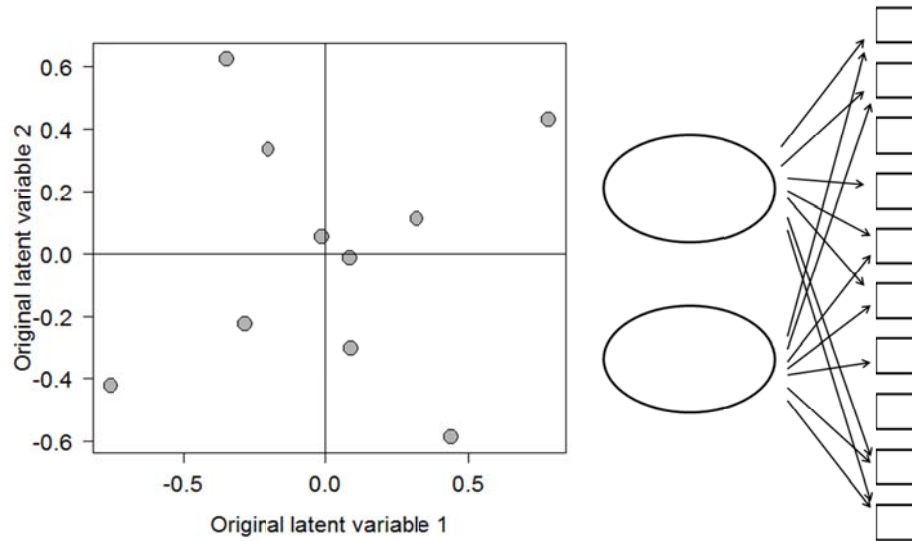
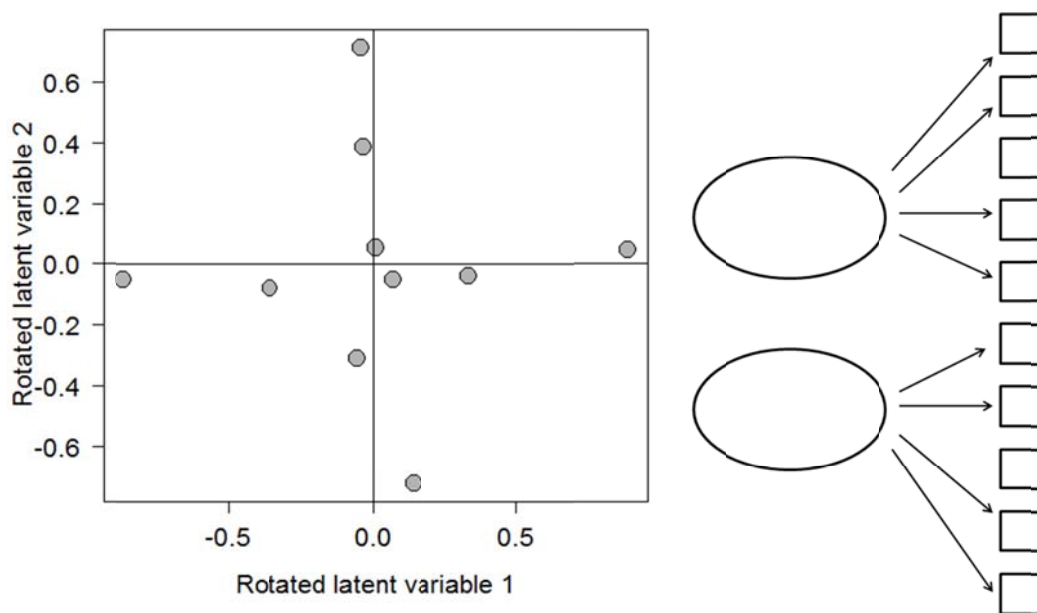


Figure 3. A scree plot from exploratory factor analysis plotted on top of a geological scree (in Austria, from Nova Development Corporation, 2004). The scree plot appears to show 2 latent variables (i.e., 2 factors) are appropriate.



Figures 4a and 4b. The loadings for an unrotated two-factor model represented in a scatter plot (Figure 4a) and a path diagram (Figure 4b). There are 14 paths shown in the path diagram (not including the item specific errors, not shown here).



Figures 5a and 5b. The loadings for the rotated two-factor model represented in a scatter plot (Figure 5a) and a path diagram (Figure 5b). There are only 8 paths shown in the path diagram (not including the item specific errors, not shown here).

Correlations

		BFNE _01	BFNE _05	BFNE _06	BFNE _08	BFNE _11	SIAS _11	SIAS _13	SIAS _14	SIAS _15	SIAS _17	SIAS _18
BFNE _01	Correlation	1	.540**	.457**	.628**	.691**	.575**	.415**	.405**	.283*	.428**	.368**
	Sig. (2-tailed)		.000	.000	.000	.000	.000	.000	.001	.019	.000	.002
	N	68	68	68	68	68	68	68	68	68	68	68
BFNE _05	Correlation	.540**	1	.641**	.747**	.602**	.598**	.606**	.517**	.408**	.433**	.485**
	Sig. (2-tailed)	.000		.000	.000	.000	.000	.000	.000	.001	.000	.000
	N	68	68	68	68	68	68	68	68	68	68	68
BFNE _06	Correlation	.457**	.641**	1	.616**	.502**	.597**	.481**	.430**	.394**	.438**	.481**
	Sig. (2-tailed)	.000	.000		.000	.000	.000	.000	.000	.001	.000	.000
	N	68	68	68	68	68	68	68	68	68	68	68
BFNE _08	Correlation	.628**	.747**	.616**	1	.678**	.516**	.405**	.444**	.248*	.500**	.391**
	Sig. (2-tailed)	.000	.000	.000		.000	.000	.001	.000	.042	.000	.001
	N	68	68	68	68	68	68	68	68	68	68	68
BFNE _11	Correlation	.691**	.602**	.502**	.678**	1	.558**	.495**	.464**	.378**	.412**	.508**
	Sig. (2-tailed)	.000	.000	.000	.000		.000	.000	.000	.001	.000	.000
	N	68	68	68	68	68	68	68	68	68	68	68
SIAS _11	Correlation	.575**	.598**	.597**	.516**	.558**	1	.540**	.517**	.510**	.434**	.525**
	Sig. (2-tailed)	.000	.000	.000	.000	.000		.000	.000	.000	.000	.000
	N	68	68	68	68	68	68	68	68	68	68	68
SIAS _13	Correlation	.415**	.606**	.481**	.405**	.495**	.540**	1	.667**	.569**	.377**	.568**
	Sig. (2-tailed)	.000	.000	.000	.001	.000	.000		.000	.000	.002	.000
	N	68	68	68	68	68	68	68	68	68	68	68
SIAS _14	Correlation	.405**	.517**	.430**	.444**	.464**	.517**	.667**	1	.685**	.661**	.715**
	Sig. (2-tailed)	.001	.000	.000	.000	.000	.000	.000		.000	.000	.000
	N	68	68	68	68	68	68	68	68	68	68	68
SIAS _15	Correlation	.283*	.408**	.394**	.248*	.378**	.510**	.569**	.685**	1	.672**	.767**
	Sig. (2-tailed)	.019	.001	.001	.042	.001	.000	.000	.000		.000	.000
	N	68	68	68	68	68	68	68	68	68	68	68
SIAS _17	Correlation	.428**	.433**	.438**	.500**	.412**	.434**	.377**	.661**	.672**	1	.710**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.002	.000	.000		.000
	N	68	68	68	68	68	68	68	68	68	68	68
SIAS _18	Correlation	.368**	.485**	.481**	.391**	.508**	.525**	.568**	.715**	.767**	.710**	1
	Sig. (2-tailed)	.002	.000	.000	.001	.000	.000	.000	.000	.000	.000	
	N	68	68	68	68	68	68	68	68	68	68	68

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Figure 6. The correlation matrix from the Social Anxiety data. This was slightly edited in the SPSS table procedure, printed, and then it was drawn on by the authors.

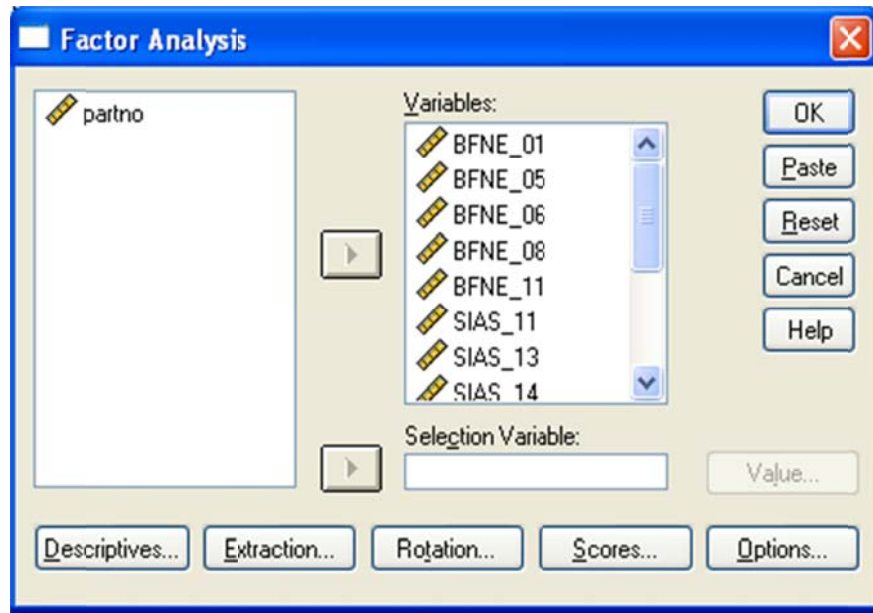


Figure 7. The SPSS dialogue box for EFA. The location of the buttons and which features are included will differ between versions (and for PASW).

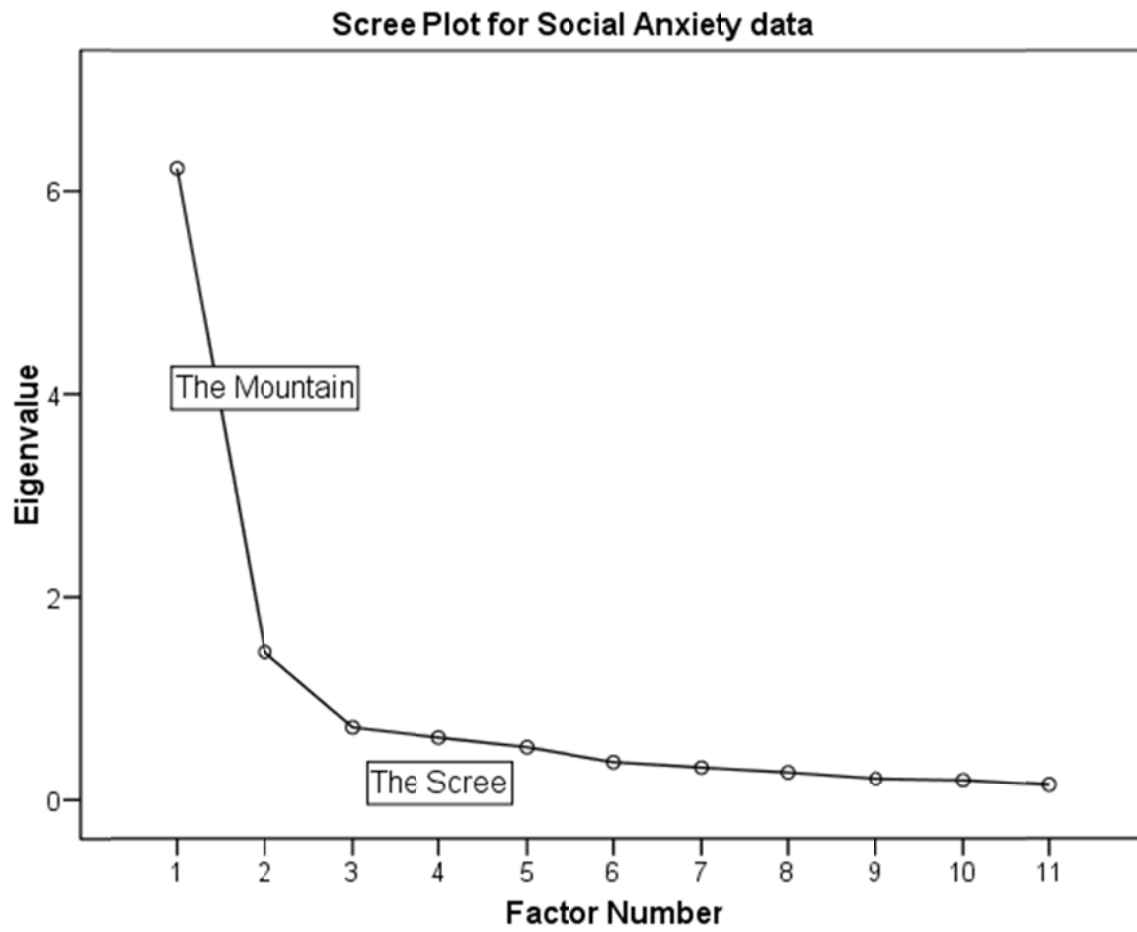


Figure 8. A scree plot for the Social Anxiety data.

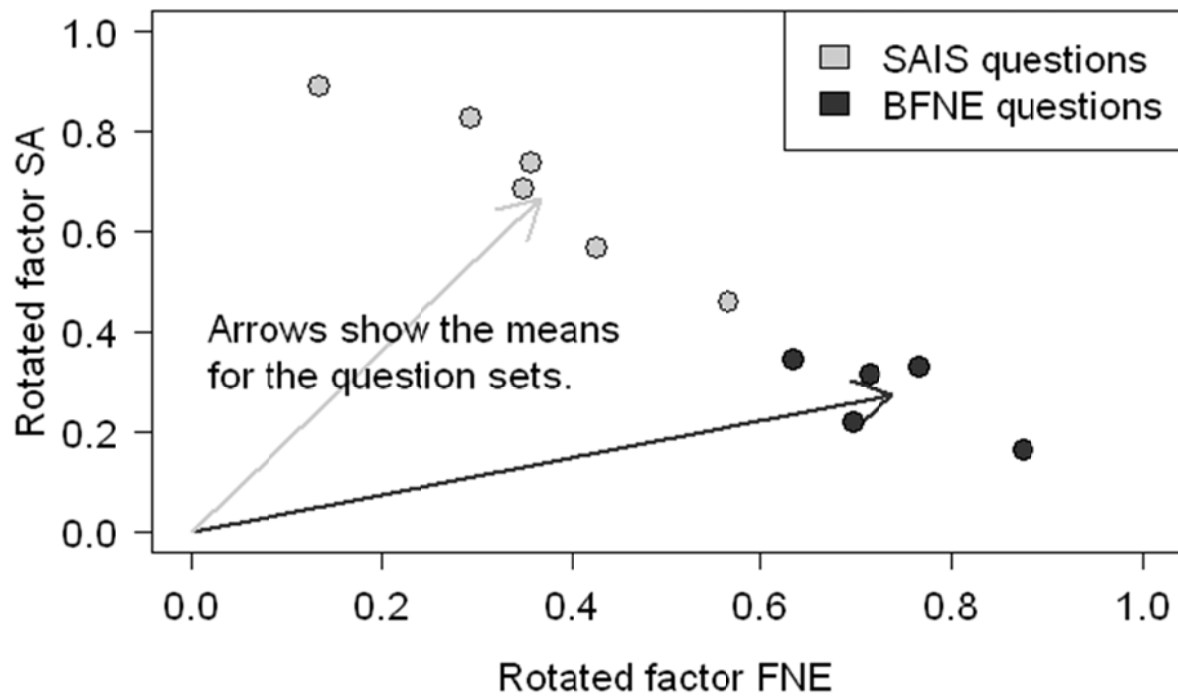


Figure 9. A plot of the rotated factor loadings for the Social Anxiety data.